MULTI-SCALE CROPPING MECHANISM FOR REMOTE SENSING IMAGE CAPTIONING

Xueting Zhang[1], Qi Wang[1], Shangdong Chen[2], Xuelong Li[1]*

[1]School of Computer Science and Center for OPTical IMagery Analysis and Learning(OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, Shaanxi, P. R. China.
[2]School of Information Science and Technology,
Northwest University, Xi'an 710072, Shaanxi, P. R. China.

ABSTRACT

With the rapid development of artificial satellite, a large number of high resolution remote sensing images can be easily obtained now. Recently, remote sensing image captioning, which aims to generate accurate and concise descriptive sentences for remote sensing images, has been promoted by template-based model and encoder-decoder model with several related datasets released. Based on an encoder-decoder model, we propose a training mechanism of multi-scale cropping for remote sensing image captioning in this paper, which can extract more fine-grained information from remote sensing images and enhance the generalization performance of the base model. The experimental results on two datasets UCM-captions and Sydney-captions demonstrate that the proposed approach availably improves the performances in describing high resolution remote sensing images.

Index Terms— Remote sensing image, image captioning, encoder-decoder, multi-scale cropping

1. INTRODUCTION

Benefited from the rapid development of deep learning, recently, many researches in the field of remote sensing have been greatly promoted, including object detection, scene classification [1], semantic segmentation and so on. These tasks mainly explore the attributes of visual features from the remote sensing images such as category, width, height and others. For the task of remote sensing image captioning, however, it has to further dig the relationship between these attributes and describe it with flexible sentences in human language. To make the most of the information in remote sensing images, remote sensing image captioning is able to extract both the visual and the context features from an image, which can help people intuitively understand the content of remote sensing images on a semantic level. It has a wide range of application prospects in many domains such as image retrieval, resource investigation, disaster detection and military intelligence generation.

Image captioning [2] is a comprehensive task which combines computer vision and natural language processing. Since that encoder-decoder based method [3] can automaticallty learn the hign-level semantic features and dig their textual relationships, it has dominated the field of image captioning with the best perpormance. The encoder process aims to represent an image with a feature map/vector by using *Convolutional Neural Networks* (CNNs), while in decoder process the feature map/vector is decoded into a sentence by a sequence model, such as *Recurrent Neural Networks* (RNNs) and *Long-Short Term Memory networks* (LSTM) [4]. This type of methods usually learn a large amount of embedded space for images and captions. Because there are no strict grammatical constraints, the system can generate relatively new textual descriptions for input images.

Correspondingly, some researches recently have been studied in remote sensing image captioning. Qu et al. [5] firstly propose a deep multi-modal neural network model to solve the problem of understanding HSR remote sensing images at the semantic level, then Shi et al. [6] present a remote sensing image captioning framework by leveraging the recent techniques of deep learning and fully convolutional networks. For the former, different CNN architectures with RNN or LSTM are combined to generate meaningful sentences for remote sensing images. While for the latter, a template-based model is applied to generate concise descriptions. Furthermore, based on a multimodal model, attention mechanism is introduced by Lu et al [7]. And both the handcrafted and deep features are utilized in this method.

However, there are many problems need to be solved. For example, it is quite significant to explore how to reduce the overfitting problem of the model with the limit of training samples. In the field of image classification, ten crops mechanism [8] is introduced to reduce the overfitting problem in the training stage, which is an effective method for data argumentation. Besides, multi-scale training [9] is a commonly

*Corresponding Author.

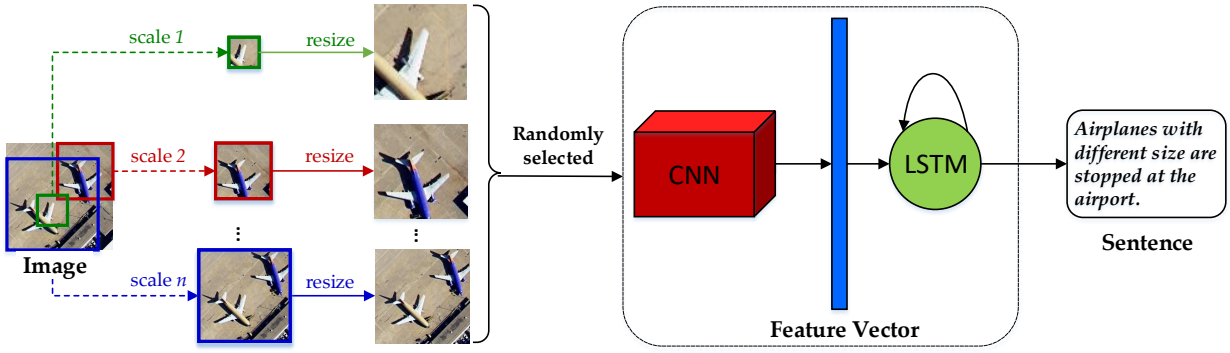

Fig. 1. The framework of the proposed method.

used method in the field of object detection. Inspired by these methods, we propose a multi-scale cropping mechanism for remote sensing image caption generation, combining ten crops with multi-scale training. The overall framework with the proposed method is shown in Fig. 1. The main contributions of this paper are summarized as follows:

1. A significant training mechanism of multi-scale cropping for remote sensing image caption generation is proposed in this paper. The proposed training mechanism can improve the performance of image captioning with the effect of reducing the overfitting problem.
2. Based on the encoder-decoder framework, we train and test different combined models of CNNs and LSTM on two datasets UCM-captions and Sydney-captions.
3. The abundant expermential results prove the effectiveness of the proposed apporach for remote sensing image captioning.

2. METHODOLOGY

The overview of whole framework is shown in Fig. 1. And the structure of the proposed method mainly contains two parts: a multi-scale cropping mechanism and an encoder-decoder based framework. Before being input into an encoder-decoder framework, the given images need to be processed by a multi-scale cropping mechanism.

2.1. Multi-scale Cropping

Noting that the effective feature extraction is critical to the task of remote sensing image captioning, we introduce a multi-scale cropping training mechanism to improve the generalization of feature representation.

As shown in Fig. 1, the input of the framework is an image that is resized to d x d. We need to select an patch of d' x d' from the image through the method of multi-scale cropping. Firstly, scale s is randomly picked up from a scale list S that is set manually in advance. And all the elements of S are no larger than 1.0. The mathematical relationship of d, d' and s is:

$$d' = d * s \tag{1}$$

In this paper, d is 256 and S is set as [1.0, 0.875, 0.66]. Thus the range of d' is [256, 224, 169]. Obviously, d' is no larger than d. Motivated by Tencrop, the patch is randomly cropped from the five corners (upper left, lower left, upper right, lower right and the center) and the flipped five corners of the images. For each image, there are 22 (10 x 2 + 2) possible cropping paths. Then the cropped patch with various size is resized to 224 x 224 before being sent to the encoder-decoder model. Thus the cropping mechanism can provide various cropped patches P_i ($i = 1, ..., 10$) from each training image, which can improve the generalization of the model.

2.2. Encoder-Decoder Framework

To generate more novel sentences for remote sensing images, our approach is mainly based on the popular encoder-decoder framework, which is divided into two stages: image representation and sentence generation. In more detail, the encoder process encodes an image into a fixed-length feature vector, while the decoder process aims to decode the feature vector into a meaningful sentence.

2.2.1. Image Reresentation

With the explosive development of deep learning and computer vision, extracting features from images by learning based method has gradually gained popularity [10]. Due to the excellent ability of automatically extracting advanced features with fewer parameters, *Convolutional Neural Networks* (CNNs) are introduced here for image representation. By replacing the last fully connected layer, three deep CNNs are

Table 1. PERFORMANCES OF THE MULTI-SCALE CROPPING MECHANISM FOR REMOTE SENSING IMAGE CAPTIONING ON THE UCM-CAPTIONS DATASET. B-N IS BLEU SCORE FOR N-GRAM.

Model	Scale	**B-1**	**B-2**	**B-3**	**B-4**
VGG-16	$[s_1]$	57.1	50.5	44.6	38.3
	$[s_1, s_2]$	58.0	50.7	45.2	39.5
	$[s_1, s_2, s_3]$	59.4	51.4	46.3	41.6
Inception-ResNetV2	$[s_1]$	54.5	46.9	41.8	36.4
	$[s_1, s_2]$	54.8	48.5	43.3	37.8
	$[s_1, s_2, s_3]$	56.7	49.7	44.2	38.8
ResNet-152	$[s_1]$	58.7	52.3	47.1	42.1
	$[s_1, s_2]$	59.1	52.6	47.1	42.4
	$[s_1, s_2, s_3]$	**59.4**	**53.2**	**48.1**	**42.9**

Table 2. PERFORMANCES OF THE MULTI-SCALE CROPPING MECHANIASM FOR REMOTE SENSING IMAGE CAPTIONING ON THE SYDNEY-CAPTIONS DATASET. B-N IS BLEU SCORE FOR N-GRAM.

Model	Scale	**B-1**	**B-2**	**B-3**	**B-4**
VGG-16	$[s_1]$	54.9	45.0	39.3	31.9
	$[s_1, s_2]$	56.3	48.1	41.9	33.7
	$[s_1, s_2, s_3]$	57.6	49.4	42.5	35.8
Inception-ResNetV2	$[s_1]$	58.2	51.6	45.5	39.4
	$[s_1, s_2]$	59.7	52.3	46.9	41.7
	$[s_1, s_2, s_3]$	60.9	53.4	47.3	41.7
ResNet-152	$[s_1]$	58.8	51.5	44.7	38.2
	$[s_1, s_2]$	60.5	52.3	45.1	38.3
	$[s_1, s_2, s_3]$	**61.5**	**54.0**	**47.3**	**40.0**

used to represent the cropped images with a fixed-length vector, which are pre-trained on ImageNet dataset. Considering the images preprocessed by the multi-scale cropping mechanism have been randomly cropped into different sizes, both the global and local features can be obtained by the CNN.

$$v_0 = \mathrm{CNN}(P_i) \tag{2}$$

As shown in formula (2), each patch P_i is transfered into CNN, and then a fixed-length feature vector v_0 is extracted from it.

2.2.2. Sentences Generation

To generate accurate descriptive sentences for remote sensing images, *Long-Short Term Memory networks* (LSTM) [4] is applied in the captioning decoding stage. By using gates to control the transmission of network information, LSTM is designed to solve the long-term dependency problem in RNN. The core of LSTM is three gates, including forgotten gate, input gate, and output gate. LSTM first decides which information to discard through the forgotten gate. Then based on the input gate, it determines the values we are going to update. Meanwhile, the tanh layer is used to generate candidate values that can be added to the network state. After that, the output layer can get the final output state by filtering.

In the time t=1, the feature vector v_0 is transferred to LSTM. When the t-th word is generated, we can represent the process as follows:

$$s = \{w_1, ..., w_t, ..., w_N\}, t \in \{0...N\} \tag{3}$$

$$h_t = g(h_{t-1}, v_0, w_{t-1}) \tag{4}$$

$$p_t = softmax(h_t) \tag{5}$$

where h_t represents the hidden state of LSTM at time t, and w_t is the corresponding word in the caption s. Specially, $g(\cdot)$ denotes the process of LSTM. After going through a softmax fuction, we can get the probability of the next word appearing i.e. p_t. The final goal of this step is to minimize the negative likelihood function of target sentences, namely $Loss$.

$$Loss = -\sum_{t=1}^{N} \log p_t(w_t) \tag{6}$$

3. EXPERIMENTS

In this section, we first give an introduction to the datasets and the metrics for experiments. Then some details of the experiments setup are given. Finally, the performance of experiments with different CNNs on two datasets are compared to verify the generalization capabilities of our method.

3.1. Datasets and Metrics

In experiments, two datasets of remote sensing image captioning, UCM-captions and Sydney-captions, are used to evaluate the performance of different architectures with the different cropping scales. The UCM-captions dataset is provided by [5], which consists of 21 classes with 100 iamges for each class. Each image is 256 $\times$ 256 pixels and the resolution is 0.3048m. Based on the Sydney Data Set, the Sydney-captions is also proposed in [5], which contains 2329 images with 7 classes. For both two datasets, five sentences are given to describe each image.

In addition, a commonly used metric for image captioning $i.e.$ BLEU [11] is utilized to evaluate the quality of the generated captions. As a measure of similarity based on accuracy, it is first proposed in [11] to evaluate the task of machine translation. In detail, it is able to analyze the consistency between n-gram occurrences in candidate and reference sentences.

3.2. Setup

Based on the encoder-decoder framework, we train and test our method with three CNN architectures in the encoding

(a) A white airplane is stopped at the airport.

(b) Lots of houses arranged neatly and a road goes through them.

(c) The waves slapping a white sand beach over and over again.

(d) Green plants flourish on both banks of the river.

Fig. 2. The results of test images and coresponding generated captions.

stage, including VGG-16, Inception-ResNetV2 and ResNet-152. In the decoding process, an LSTM is used to generate the effective sentences. As for LSTM, the dimension of word embedding and hidden state are set to 512 and 512, respectively. And the number of layers in LSTM is one layer. We use Stochastic Gradient Descent (SGD) with the learning rate of 0.0001 to optimize the whole model.

3.3. Results and Analysis

Table 1 and Table 2 show the experimental results of the two datasets with different CNN architectures, respectively. Noting that s_1, s_2, s_3 respectively denote the scale of 1.0, 0.875 and 0.66. From the results we can find that the evaluation scores on both UCM-captions and Sydney-captions with all the CNN architectures are getting higher, with the selected scales increasing. And the Resnet-152 is the best CNN architecture for encoding process between all the architectures in experiments. The abundant experimental results prove the effectiveness of the proposed Multi-Scale Cropping Mechanism.

4. CONCLUSION

In this paper, we propose a multi-scale cropping mechanism for training, which can extract advanced semantic features from images so that it can generate meaningful sentences for the task of remote sensing image captioning. Based on an popular encoder-decoder framework, three CNNs combined with LSTM are compared to validate the generalization performance of our method. And the experimental results on two datasets show the effectiveness of the proposed method.

5. ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant U1864204 and 61773316, Natural Science Foundation of Shaanxi Province under Grant 2018KJXX-024, and Project of Special Zone for National Defense Science and Technology Innovation.

6. REFERENCES

[1] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, , no. 99, pp. 1–13, 2018.

[2] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *International Conference on Computer, Information and Telecommunication Systems*, 2016, pp. 124–128.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[4] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[5] B. Qu, X. Li, D. Tao, and X. Lu, "Deep semantic understanding of high resolution remote sensing image," in *International Conference on Computer, Information and Telecommunication Systems*, 2016, pp. 124–128.

[6] Z. Shi and Z. Zou, "Can a machine generate humanlike language descriptions for a remote sensing image?," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3623–3634, 2017.

[7] X. Lu, B. Wang, X. Zheng, and X. Li, "Exploring models and data for remote sensing image caption generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 2183–2195, 2018.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *International Conference on Neural Information Processing Systems*, 2012, pp. 1097–1105.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[10] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced NetVLAD with and weighted triplet loss for place recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2019.

[11] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Association for Computational Linguistics*, 2002, pp. 311–318.